Efficient Bootstrap Resampling

Radu V. Craiu
Department of Statistics, The University of Chicago

This paper is dedicated to my father, *Professor Virgil Craiu*

Abstract

Bootstrap principle is briefly reviewed. Hall's (1989) antithetic variates method for bootstrap is discussed and extended to more than two antithetic resampling processes. We illustrate the theory with a simulation study. The numerical results show that increasing the number of antithetic resampling processes produces significant smaller variances of the bootstrap estimator over the paired case.

Key words: Antithetic variates; Bootstrap; Efficiency; Monte Carlo

1 Bootstrap method

The bootstrap method has been known to statisticians for a long time before Efron (1979) found a suggestive name for it and emphasized that its scope is much broader than previously thought. The name *bootstrap* is a reference to the old saying according to which, one can pull oneself out of trouble by pulling one's bootstraps. Although extremely suggestive, the name is somewhat misleading in that it conveys the impression that bootstrap builds "something from nothing" while, in fact, the technique has a sound theoretical foundation.

In its basic form, the idea is quite natural. Let's assume that we want to estimate a characteristic of a population distribution F, e.g. the mean of F

$$\mu = \int x dF(x). \tag{1}$$

If we dispose of a sample of size n from F, $\mathbf{x} = \{x_1, \ldots, x_n\}$ it is natural to replace F in equation (1) with its closest available discrete counterpart, $\hat{F}$, the empirical distribution function associated to the sample $\mathbf{x}$, and estimate μ with

$$\bar{X} = \int x d\hat{F}(x). \tag{2}$$

It is worth emphasizing that the above approach is not practicable for all functionals. Moreover, in most applications the bootstrap statistics are hard to compute. Efron showed that a way around this difficulty is the use of Monte Carlo methods, specifically, "bootstrap resampling". In what follows, a *resample* will be, given the sample $\mathbf{x}$, an unordered collection of n items $\mathbf{x}^* = \{x_1^*, \ldots, x_n^*\}$ sampled with replacement from $\mathbf{x}$. We will denote by F^* the empirical distribution function of $\mathbf{x}^*$.

In most cases where statistical inference is needed, we try to find a way to describe the relationship that exists between a sample and the population from which the sample has been drawn. Hall (1992) formally states the problem as follows: given a class of functionals $\{f_t : t \in \mathcal{T}\}$ we need to find t_0 such that f_{t_0} is the solution of the equation

$$E[f_t(F, \hat{F})|F] = 0. \tag{3}$$

For example, if

$$\mu_r = \left[\int x dF(x)\right]^r$$

is the r-th power of the population distribution mean, then the sample estimate will be

$$\hat{\mu}_r = \left[\int x d\hat{F}(x)\right]^r$$

To correct for bias one would like to find t_0 the solution of

$$E[f_t(F, \hat{F})|F] = E[\mu_r - \hat{\mu}_r + t|F] = 0.$$

Our bias corrected estimate for μ_r will be then $\hat{\mu}_r + t_0$.

To obtain an approximate solution to equation (3) we will apply our assumption that the relationship between F and $\hat{F}$ is well approximated by the relationship

existent between $\hat{F}$ and F^*. An approximate solution to (3) will then be the solution to

$$E[f_t(\hat{F}, F^*)|\hat{F}] = 0 \tag{4}$$

In most cases the conditional expectation in (4) is approximated using samples from a distribution determined by the original sample $\mathbf{x}$.

We would like to emphasize that the bootstrap method is not needed in all circumstances. The underlying idea is that of replacing the true distribution function F with $\hat{F}$ and $\hat{F}$ with F^* in a formula that expresses a parameter as a functional of F. For applications where the substitutions are unlikely to produce estimates similar to the originals, the method will produce unreliable results.

2 Antithetic variates

The word *antithetic* refers to the main objective of the method, that is, to produce random numbers that are *negatively correlated*. The reason for which statisticians are interested in such negatively correlated numbers is emphasized by the following simple example. Consider the problem of estimating the integral

$$\xi = \int_a^b h(x)dx$$

by Monte Carlo. The standard crude Monte Carlo estimator uses a sample drawn uniformly on (a, b), $x_1, ..., x_n$ and approximates ξ with $(b-a)\sum_{i=1}^n h(x_i)/n$. The antithetic principle (Hammersley and Morton, 1956) states that the above estimate will be subject to less variability if, for each x_i, we also use its "mirror" $x_i' = a + (b - x_i)$. This mirror variate is called an antithetic variate and its use can be effective in reducing the variance of the Monte Carlo estimate. For a sample of size n, one can combine the sampled points and their mirrored counterparts into

$$\frac{b-a}{n}\sum_{i=1}^{n}(h(x_i) + h(x_i')).$$

The key ingredient of the method is the negative correlation induced between $h(X)$ and $h(X')$. It is natural then to suppose that the use of antithetic variates is related to a certain monotonicity structure existent in the problem (in the above simple example, h should be monotone on (a, b) for the variance reduction to surely take place).

It is known that the Fréchet-Hoeffding inequality's lower bound (Fréchet, 1951) for two-dimensional distribution functions is a distribution function itself (e.g. Joe, 1998). This makes $(U, 1-U)$, where $U \sim U(0,1)$, the best choice for a pair of antithetic variates (see also Whitt, 1976). Unfortunately, for a number of random variates larger than two, the lower bound is no longer a distribution function and an *uniformly best* solution to the problem is unknown. Craiu and Meng (2001) are proposing a few methods to generate K antithetic variates and discuss ways to implement the antithetic principle in approximate and exact Markov chain Monte Carlo (MCMC). For good introductory references to exact MCMC algorithms see Casella, Lavine, and Robert (1999) and Wilson (2000). In the next section we will discuss the antithetic principle for bootstrap as presented by Hall (1989) and we will perform a numerical experiment that will show that the increase in the number of antithetic processes results in significant efficiency gains. We will take a different approach then the one recommended for MCMC algorithms since, for bootstrap, the method aims at antithetically sampling from discrete distributions with a finite support.

3 Antithetic resampling for the bootstrap

In the present section we present an antithetic variates method for conducting bootstrap resampling operations. We will follow in the tracks of Hall (1989) but we will also extend his method to more than two parallel resampling operations at a time.

Suppose we are interested in the expected value μ of an order-invariant statistic

$\theta(X_1, \ldots, X_n)$. The bootstrap estimate of this quantity is

$$\hat{\mu} = E[\theta(X_1^*, \ldots, X_n^*)|\mathbf{X}]$$

where $\mathbf{X} = \{X_1, \ldots, X_n\}$ is the random sample and $\{X_1^*, \ldots, X_n^*\}$ are drawn with replacement from $\mathbf{X}$. In practice, we approximate $\hat{\mu}$ with

$$\hat{\mu}^* = \frac{1}{B}\sum_{b=1}^{B} \theta(X_{b1}^*, \ldots, X_{bn}^*) \tag{5}$$

where, conditional on $\mathbf{X}$, X_{bi}^* are independently and uniformly distributed on $\mathbf{X}$. The samples X_i are d-dimensional, with $d \geq 1$. Another way of writing (5) is

$$\hat{\mu}^* = \frac{1}{B}\sum_{b=1}^{B} \theta(X_{I(b,1)}^*, \ldots, X_{I(b,n)}^*)$$

where $I(b, i)$ are independently and uniformly distributed on the integers $1, \ldots, n$.

Antithetic resampling is based on antithetic permutations $\pi_1, \ldots, \pi_k$ of the integers $1, \ldots, n$. If the π's are chosen appropriately and

$$\mu_j^* = \frac{1}{B}\sum_{b=1}^{B} \theta(X_{\pi_j(I(b,1))}^*, \ldots, X_{\pi_j(I(b,n))}^*),$$

then μ_j^*, $j = 1, \ldots, k$ should be negatively correlated conditional on $\mathbf{X}$.

To appreciate the form the antithetic permutations should take we will assume that $\theta(x_1, \ldots, x_n)$ is actually a function of the mean $\frac{1}{n}\sum_{i=1}^n x_i$. We will write $\theta(x_1, \ldots, x_n) = \theta(\frac{1}{n}\sum_{i=1}^n x_i)$.

Assume that the vectors X_i are d-variate and θ is smooth and define $\theta_i(x) = \frac{\partial}{\partial x^i}\theta(x)$ and

$$Y_h = \sum_{i=1}^{d} \theta_i(\bar{X})(X_h - \bar{X})^i$$

where the superscript i on a vector denotes the i-th element of that vector and $\bar{X} = \frac{1}{n}\sum_i X_i$

We relabel the sample values X_i so that $Y_1 \leq Y_2 \leq \ldots \leq Y_n$.

By Taylor expansion we have

$$\hat{\mu}_j^* = B^{-1}\sum_{b=1}^{B}\theta(\bar{X}_{b,\pi_j}^*) = \theta(\bar{X}) + B^{-1}\sum_{b=1}^{B}\sum_{i=1}^{d}\theta_i(\bar{X})(\bar{X}_{b,\pi_j}^* - \bar{X})^i + \dots$$

where $\bar{X}_{b,\pi_j}^* = \frac{1}{n}\sum_{i=1}^{n} X_{\pi_j(I(b,i))}$.

Following Hall (1989) we have that

$$Var(\hat{\mu}_j^*) = (Bn)^{-1}n^{-1}\sum_{i=1}^{n} Y_i^2$$

and if $g \neq h$

$$cov(\hat{\mu}_h^*, \hat{\mu}_g^*) = (Bn)^{-1}cov(Y_{\pi_h(I(b,n))}, Y_{\pi_g(I(b,n))}|s) = (Bn)^{-1}n^{-1}\sum_{i=1}^{n} Y_{\pi_h(i)}Y_{\pi_g(i)}$$

where all errors of approximation are $O(B^{-1}n^{-2})$.

Therefore, if we want to use k antithetic resampling processes , we need to find $k-1$ permutations $\pi_2, ..., \pi_k$ such that for any $Y_1 \leq Y_2 \dots \leq Y_n$ we have, for π_1 the identical permutation

$$\sum_{g\neq h}\sum_{i=1}^{n} Y_{\pi_g(i)}Y_{\pi_h(i)} = \text{ minimum possible} \tag{6}$$

If $k = 2$ an uniform optimal solution is $\pi_1(i) = i$ and $\pi_2(i) = n - i + 1$ for all $1 \leq i \leq n$. The solution is uniform optimal in the sense that for any $Y_1 \leq Y_2 \leq \dots \leq Y_n$ the sum $\sum_j Y_j Y_{\pi_1(j)}$ is minimal.

For $k \geq 2$ such an optimal solution doesn't exist. For instance, if $k = 3$, for each arbitrarily fixed sequence $Y_1 \leq Y_2 \leq \dots \leq Y_n$ one can find π_1, π_2, π_3 such that $\sum_{j=1}^{n}(Y_{\pi_1(j)}Y_{\pi_2(j)} + Y_{\pi_2(j)}Y_{\pi_3(k)} + Y_{\pi_1(j)}Y_{\pi_3(k)})$ is minimum possible, but the permutations may change as a new sequence is drawn. Although an algorithm to determine π_1, π_2, and π_3 for each sequence of Y's can be constructed, such an endeavor is expensive in terms of computing effort. Instead, we propose a way to combine variance reduction improvement and reduced additional computer effort by devising an algorithm that remains unchanged once k is fixed.

One can show that if we think of the solution as unique (although it isn't!) then the optimal solution is such that $\pi_1(i) = i$ and π_2, π_3 have to be permutations obtained as products of 3–cycles. We will use in our algorithm this information although we know it is not the optimal choice for any vector of Y's. We choose the entries for each cycle such that they are situated "symmetrically" away from the extremities of the sequence $Y_1 \leq Y_2 \leq ... \leq Y_n$. Of course, since we are dealing with cycles of odd length, the term "symmetric" is loosely used here. The 3–cycles we choose to use are among those that minimize the sum of the form (6) for $n = 3$ and when the Y's are equidistant: **(I)**: (1 3 2), and **(II)**: (1 2 3). The way we mimic the "grouping of the extremes" that takes place in the case $k = 2$ is, for $k = 3$, to include in the 3-cycles numbers that are at the opposite ends of the sequence $Y_1 \leq Y_2 \leq ... \leq Y_n$. Specifically, the first cycle will contain $\{1, n, n-1\}$, the second, $\{2, 3, n-2\}$ and so on until there are less than three entries from $\{1, 2, \ldots, n\}$ that are not included in one of the cycles. Those elements, if any, will be left unchanged (if there is only one left) or will be transposed by one of the permutations. As a result, we define π_2 as the product of cycles of the type **(I)**, and π_3 as the product of cycles of the type **(II)** and each the first 3–cycle from π_2 and , respectively, π_3, will contain the same entries. For example, if $n = 7$, π_2 and π_3 will be

$$\pi_2 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 7 & 5 & 2 & 4 & 3 & 1 & 6 \end{pmatrix}$$

and

$$\pi_3 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 6 & 3 & 5 & 4 & 2 & 7 & 1 \end{pmatrix}.$$

Although the scheme is not uniformly optimal, simulations show that important variance reductions take place when we use π_1, π_2, and π_3 to increase the number of antithetic resampling processes to $k = 3$. We emphasize that one can write a subroutine which generates π_2, π_3 automatically once n is given since the construction depends only on n *mod* 3.

In the following table we summarize the variance reductions obtained when we compute bootstrap estimates of a mean $E[X] + E[Y]$ based on a i.i.d. sample of size 50 from $(X \sim N(3, \sigma^2), Y \sim Gamma(1, 1))$. Following the recommendations given in Efron and Tibshirani (1994) we used 210 bootstrap resamples and we estimated the variance reduction using a Monte Carlo sample of size 5000. The entry in a given cell represents the ratio between the Monte Carlo estimate of the bootstrap estimator's variance when using k antithetic processes and the same variance computed with independent bootstrap resamples. We would like to stress that the technique is applicable to other problems, like estimation of cumulative distribution functions (see also Hall, 1989 for more illustrations).

$k \setminus \sigma$	1	2	3	4	6	8	10	20
2	0.89	0.90	0.78	0.65	0.65	0.68	0.71	0.68
3	0.71	0.39	0.38	0.32	0.33	0.32	0.36	0.27

Table 1: Variance reduction ratio Var_{anti}/Var_{indep} for different values of σ and k antithetic resampling processes

It is striking that when σ is large and compensates for the skewness in the Gamma distribution, the paired antithetic variates are doing well but by using $k = 3$ we shrink the variance reduction ratio by more than 50% relative to the $k = 2$ case. Moreover, while for $k = 2$ the variance reduction is smaller as the population variance gets smaller, the pattern doesn't appear as obvious in the $k = 3$ case. We are left with the impression that the increase in efficiency is more robust when a higher number of processes are run in parallel. More theoretical work needs to be done to explain this pattern.

Therefore, if we were to leave the reader with a single message, this would contain the advice that whenever the antithetic principle is recommended, one should

implement it with more than two antithetically paired processes.

Acknowledgments

The author thanks Professor Xiao-Li Meng for encouragement and helpful discussions.

References

[1] George Casella, Michael Lavine and Christian Robert. Explaining the Perfect Sampler. *Technical Report*, ISDS, Duke University, Durham, 2000.

[2] Radu V. Craiu and Xiao-Li Meng. Multi-process parallel antithetic coupling for forward and backward Markov chain Monte Carlo. *Technical Report*, Department of Statistics, The University of Chicago, 2001.

[3] Radu V. Craiu. Multivalent Framework for Approximate and Exact Sampling and Resampling. *Doctoral Dissertation*, Department of Statistics, The University of Chicago, 2001.

[4] Bradley Efron and Rob Tibshirani. *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1994.

[5] Maurice Fréchet. Sur les tableaux de corrélation dont les marges sont données. (French) *Ann. Univ. Lyon. Sect. A.* 14: 53–77, 1951.

[6] Peter Hall. Antithetic resampling for the bootstrap. *Biometrika*, 76:713–724, 1989.

[7] Peter Hall. *The Bootstrap and Edgeworth Expansion*. Springer, New York, 1992.

[8] D. C. Hammersley and K. V. Morton. A new Monte Carlo technique: antithetic variates. *Proc. Camb. phil. Soc.*, 52:449-475, 1956.

[9] Harry Joe. *Multivariate Models and Dependence Concepts*. Chapman and Hall, New York, 1997.

[10] Wei-Liem Loh. On Latin hypercube sampling. *Ann. Stat.*, 24:2058–2080, 1996.

[11] M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245, 1979.

[12] David B. Wilson. Layered multishift coupling for use in perfect sampling algorithms (with a primer on CFTP). In Neil Madras, editor, *Monte Carlo Methods - Fields Institute Communications* vol. 26, 141–176, 2000.

[13] Ward Whitt. Bivariate distributions with given marginals. *Ann. Statist.*, 4:1280–1289, 1976.